Medical Image Analysis 94 (2024) 103153

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Diabetic foot ulcers segmentation challenge report: Benchmark and analysis

Moi Hoon Yap [a,p,*], Bill Cassidy [a], Michal Byra [b,c], Ting-yu Liao [d], Huahui Yi [e], Adrian Galdran [f,g], Yung-Han Chen [h], Raphael Brüngel [i,j,k], Sven Koitka [k,l], Christoph M. Friedrich [i,j], Yu-wen Lo [d], Ching-hui Yang [d], Kang Li [e,n], Qicheng Lao [m,n], Miguel A. González Ballester [f], Gustavo Carneiro [o], Yi-Jen Ju [h], Juinn-Dar Huang [h], Joseph M. Pappachan [p,q], Neil D. Reeves [q], Vishnu Chandrabalan [p], Darren Dancey [a], Connah Kendrick [a]

[a] Department of Computing and Mathematics, Manchester Metropolitan University, John Dalton Building, Chester Street, Manchester M1 5GD, United Kingdom
[b] Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland
[c] RIKEN Center for Brain Science, Wako, Japan
[d] Department of Computer Science, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan
[e] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
[f] BCN Medtech, Universitat Pompeu Fabra, Barcelona, Spain
[g] AIML, University of Adelaide, Australia
[h] Institute of Electronics, National Yang Ming Chiao Tung University, No. 1001, University Road, Hsinchu 300, Taiwan
[i] Department of Computer Science, University of Applied Sciences and Arts Dortmund (FH Dortmund), Emil-Figge-Str. 42, 44227 Dortmund, Germany
[j] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Zweigertstr. 37, 45130 Essen, Germany
[k] Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Girardetstr. 2, 45131 Essen, Germany
[l] Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Hufelandstr. 55, 45147 Essen, Germany
[m] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[n] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[o] University of Surrey, Guildford, United Kingdom
[p] Lancashire Teaching Hospitals NHS Trust, Preston, PR2 9HT, United Kingdom
[q] Department of Life Sciences, Manchester Metropolitan University, Manchester, M1 5GD, United Kingdom

ARTICLE INFO

Keywords:
Deep learning
Diabetic foot ulcers
Segmentation
Convolutional neural networks
Metrics

ABSTRACT

Monitoring the healing progress of diabetic foot ulcers is a challenging process. Accurate segmentation of foot ulcers can help podiatrists to quantitatively measure the size of wound regions to assist prediction of healing status. The main challenge in this field is the lack of publicly available manual delineation, which can be time consuming and laborious. Recently, methods based on deep learning have shown excellent results in automatic segmentation of medical images, however, they require large-scale datasets for training, and there is limited consensus on which methods perform the best. The 2022 Diabetic Foot Ulcers segmentation challenge was held in conjunction with the 2022 International Conference on Medical Image Computing and Computer Assisted Intervention, which sought to address these issues and stimulate progress in this research domain. A training set of 2000 images exhibiting diabetic foot ulcers was released with corresponding segmentation ground truth masks. Of the 72 (approved) requests from 47 countries, 26 teams used this data to develop fully automated systems to predict the true segmentation masks on a test set of 2000 images, with the corresponding ground truth segmentation masks kept private. Predictions from participating teams were scored and ranked according to their average Dice similarity coefficient of the ground truth masks and prediction masks. The winning team achieved a Dice of 0.7287 for diabetic foot ulcer segmentation. This challenge has now entered a live leaderboard stage where it serves as a challenging benchmark for diabetic foot ulcer segmentation.

1. Introduction

Following the successes of previous Diabetic Foot Ulcers Challenges (DFUC), i.e. DFUC 2020 (Cassidy et al., 2021b) and DFUC 2021 (Yap et al., 2021a), DFUC 2022 focused on segmentation (Kendrick et al., 2022). This paper reports on the insights of the DFUC 2022 and conducts a post-analysis of the participants' methods and results. We conduct a comprehensive analysis on the performance of the winning algorithms by studying three ensemble methods, conducting statistical

* Corresponding author.
E-mail address: m.yap@mmu.ac.uk (M.H. Yap).

https://doi.org/10.1016/j.media.2024.103153
Received 27 June 2023; Received in revised form 30 January 2024; Accepted 20 March 2024
Available online 24 March 2024

analysis on the results, analysing region-based segmentation, and investigating the relationship between Dice similarity coefficient (DSC) values and region sizes. This research has been completed in adherence to the Biomedical Image Analysis Challenges (BIAS) guidelines (Maier-Hein et al., 2020a), as approved by the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) initiative (Pandis and Fedorowicz, 2011).

2. Challenge description

The diabetic foot ulcers challenges (2020–2022) are a series of annual deep learning challenges hosted by the Medical Image Computing and Computer Assisted Interventions (MICCAI) society for the international conferences which they organise each year. The purpose of these challenges is to develop fully automated deep learning methods for localisation, classification, and segmentation of DFU. The first diabetic foot ulcer challenge (DFUC 2020) focused on DFU localisation methods for automated DFU detection. Two winning teams were declared for DFUC 2020: (1) Ryo Hachiuma, Hiroki Kajita and Hideo Saito (Keio University and Keio University School of Medicine, Japan) achieved the highest mAP (0.6940), and (2) Manu Goyal and Saeed Hassanpour (Dartmouth College, USA) achieved the highest F1-score (0.7434) (Yap et al., 2021b). The second challenge (DFUC 2021) focused on multi-class classification for 4 classes (control, infection, ischaemia, and both infection and ischaemia). The winner of DFUC 2021 was Adrian Galdran (Bournemouth University, UK) with a macro F1-score of 0.6216 (Cassidy et al., 2021a). The third challenge (DFUC 2022) focused on delineation of DFU at pixel level, which is a key task for wound area measurement.

2.1. Organisation

The diabetic foot ulcer challenges 2020–2022 were co-organised by researchers from the United Kingdom (The Manchester Metropolitan University (MMU), Lancashire Teaching Hospitals (LTH) and University of Manchester (UoM)), New Zealand (Waikato District Health Board (WDHB)), the United States (University of Southern California (USC) and Baylor College of Medicine (BCM)) and India (Manipal College of Health Professions (MCHP)). The diabetic foot ulcer challenge 2022 was organised by Moi Hoon Yap (MMU), Neil Reeves (MMU), Andrew Bolton (UoM), Satyan Rajbhandari (LTH), David Armstrong (USC), Arun G. Maiya (MCHP), Bijan Najafi (BCM), Bill Cassidy (MMU), and Justina Wu (WDHB).

The goal of the 2022 challenge was to evaluate the performance of computer algorithms in diabetic foot ulcers (DFU) segmentation.

2.2. Dataset preparation

Medical photographs of DFU wounds were acquired from diabetic patients at the Lancashire Teaching Hospitals NHS Foundation Trust by two clinical experts in podiatry. The DFU wound photographs were acquired using three digital cameras: a Kodak DX4530 (5 megapixel), a Nikon COOLPIX P100 (10.3 megapixel), and a Nikon D3300 (24.2 megapixel). All DFU wound photographs were acquired with close-ups of the patient's foot using auto-focus, with zoom or macro functions disabled. A camera aperture setting of f/2.8 was used, with photographs taken at a distance of approximately 30–40 cm with a parallel orientation to the plane of the DFU. Flashes were deactivated, with room lighting used as the primary light source. The DFU wound photographs were distributed between 5 podiatrists, each with more than 5 years of clinical experience. Instructions were provided to the experts to delineate the ulcer regions using the VGG Image Annotator software tool (Dutta et al., 2016; Dutta and Zisserman, 2019). The polygon regions defined by the experts were then smoothed using a snake active contour algorithm (Kroon, 2022), followed by conversion to binary masks, with black pixels representing the background, and white pixels representing wound regions. The binary masks were used as ground truth for both training and testing sets. The original DFU wounds photographs were captured at various resolutions, therefore, as a preprocessing stage all photographs and corresponding masks were resized to 640×480 pixels as a standardisation measure. Ethical approval was obtained from the UK National Health Service (NHS) Research Ethics Committee (REC) to use these images for the purpose of research. The NHS REC reference number is 15/NW/0539. Written informed consent was obtained from all participating patients. As in DFUC 2020, the dataset was divided into two main sets of images and corresponding binary masks — the training set ($n = 2000$) and the testing set ($n = 2000$). We divided the data evenly to ensure that models could be trained and tested sufficiently. Prior chronic wound datasets comprised relatively small test sets (approximately 20%) which may not sufficiently challenge trained models. Therefore, we determined that a 50:50 split would help towards obtaining more accurate test metrics.

2.3. Leaderboard management

The Grand-Challenge online platform (https://dfuc2022.grand-challenge.org/) was used to process three leaderboard submission phases, i.e., validation stage, testing stage, and a live leaderboard to continue to support the research community after the challenge deadline. Participants were required to submit prediction masks to the online challenge platform with pixel-wise labels for background (0) and ulcer regions (1). A paper highlighting the contribution of the submission, including the method description, experimental results and analysis, and a GitHub repository URL containing all source code was also required (in accordance to the format stipulated by MICCAI 2022). The evalutils (Meakin, 2018) software library was used to measure the performance of segmentation accuracy of participant prediction masks. During the validation stage, participants were permitted a maximum of 10 submissions per day over a period of 6 weeks. The validation stage was used for sanity checking and fine-tuning of models using the validation dataset (a subset of the testing set). During the testing stage, participants were limited to submit once per day over a period of two weeks. The results were not released during the testing stage to prevent participants overfitting their models to the testing set.

2.4. Dataset usage and participation policy

Participants were permitted to use non-challenge datasets for training and validation purposes. This included other publicly available DFU datasets. Additionally, participants were permitted to use their own datasets on the basis that those datasets were shared publicly with the research community. Participants were permitted to use the dataset for non-commercial purposes only. Additionally, participants were prohibited from modifying the ground truth masks. Organisations or companies who were affiliated with members of the organising committee were not excluded from participation in the challenge. However, such organisations/companies were required to ensure that their submissions were completely independent of the members of the organising committee.

2.5. Results announcement and award policy

All challenge results were made available publicly on the DFUC 2022 website (https://dfu-challenge.github.io/) and the Grand Challenge website (https://dfuc2022.grand-challenge.org/). The top-5 performing methods were then invited to the in-person challenge event to present their work. Certificates were provided to the top-3 performing teams. Prizes were also awarded to the top-3 performing teams, which were provided by AITIS who were the challenge sponsors. The prizes awarded were wearable monitoring sensor devices.

2.6. Challenge schedule and publication policy

The training data was released on the 27th April 2022. Following this, the validation data was released on the 21st June 2022. The test data was released on the 1st July 2022, with a final submission deadline on the 29th July 2022. The winner and invitation speakers were announced on the 15th August 2022. All challenge deadlines were subject to change according to MICCAI 2022 scheduling changes. The challenge organisers were responsible for publishing one or more challenge journal papers which reported on the challenge results. Participating authors were permitted to publish their papers separately, with decisions on publication strategy made according to achieving publication in the highest ranking journals.

2.7. Conflict of interest statement and test label safeguarding

No external funding was received in relation to the DFUC 2022. Additionally, no funding was received from the challenge sponsors (AITIS). Ground truth masks for the DFUC 2022 test set are accessible only to the following MMU Computer Vision Laboratory researchers: Moi Hoon Yap, Connah Kendrick, and Bill Cassidy.

2.8. Metrics and evaluation

To assess the performance of the algorithms developed by participants, we determine segmentation accuracy in terms of DSC, Jaccard index, False Positive Error (FPE), and False Negative Error (FNE). Image-based metrics were used to allow multiple DFU wounds to be evaluated as a single wound per image.

DSC was used to determine overall leaderboard rankings, and is defined as two times the area of the intersection of X (ground truth) and Y (prediction), divided by the sum of the areas of X and Y. DSC values are reported in the range of $0-1$, where 0 indicates no overlap, and 1 indicates a perfect overlap. DSC is denoted by Eq. (1).

$$DSC = 2 * \frac{|X \cap Y|}{|X| + |Y|} = \frac{2 * TP}{2 * TP + FP + FN} \tag{1}$$

where TP is True Positives, FP is False Positives and FN is False Negatives. In the case of ties in DSC, the Jaccard index, also known as Intersection over Union (IoU), is used as the second metric for the leaderboard rankings. IoU is defined as the area of intersection divided by the area of union, and is expressed as Eq. (2).

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} = \frac{TP}{TP + FP + FN} \tag{2}$$

The FPE indicates the ratio of a method which falsely predicts a non-DFU pixel as a DFU pixel, and is defined in Eq. (3).

$$FPE = \frac{FP}{FP + TN} \tag{3}$$

where TN is True Negatives. The FNE indicates the ratio of a method which falsely predicts a DFU pixel as non-DFU pixel, and is denoted as in Eq. (4).

$$FNE = \frac{FN}{FN + TP} \tag{4}$$

Both DSC and IoU assume that an overlap is present. In cases where prediction masks show no overlap with ground truth masks, a score of 0 is assigned.

3. Summary of challenge methods

Since the opening of DFUC 2022, we received 72 requests from 47 countries to obtain the challenge datasets. A total of 26 teams participated in the challenge. In the DFUC 2022 proceedings (Kendrick et al., 2023), we summarise the methods from the top-10 teams, who have submitted their challenge papers and presented at the DFUC 2022 conference in conjunction with MICCAI 2022, conducted in Singapore on the 22nd September 2022. This post-challenge analysis focuses on the performance of the five winning teams who achieved a DSC > 0.7. We conduct a size analysis based on the per ulcer-based segmentation, examine the effectiveness of ensemble models, and perform statistical analysis on the results. The top-5 participating teams who were eligible for inclusion in the present challenge report are as follows: (1st place) Yllab Team, (2nd place) LkRobotAILab Team, (3rd place) AGaldran Team, (4th place) ADAR-LAB Team, and (5th place) FHDO Team.

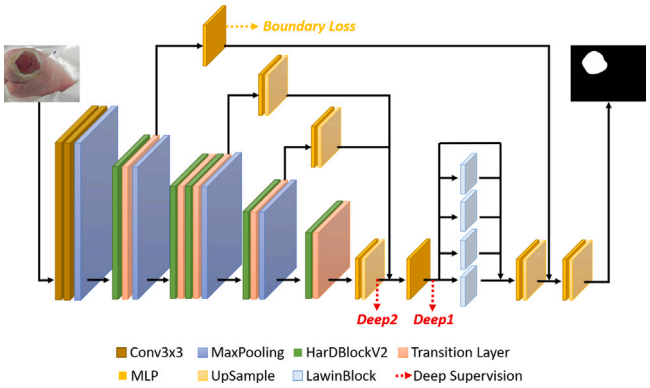

Fig. 1. Overview of HarDNet-DFUS.

3.1. Liao et al. (1st place, Yllab team)

Liao et al. (2023) proposed HarDNet-DFUS, as depicted in Fig. 1. It consists of an encoder backbone with a new HarDBlkV2 module and the decoder with a Lawin Transformer (Yan et al., 2022). The backbone of the previous state-of-the-art HarDNet-MSEG (Huang et al., 2021) (used for colonoscopic polyp segmentation) was enhanced and repurposed for DFU segmentation. HarDBlockV2 is modified from HarDBlock by referring to the concepts of CSPNet (Wang et al., 2020) and ShuffleNetV2 (Ma et al., 2018). The following three enhancements were implemented in the network design:

1. Channel splitting was performed on the convolutional layer according to its output connection number, which can reduce the DRAM access to achieve the optimal MACs over CIO ratio (MoC), as proposed by HarDNet (Chao et al., 2019b).
2. Inter-layer connectivity is performed based on the factors of the required block depth, simplifying the design of the network architecture so that the depth of the basic building block is no longer limited to a power of 2.
3. A squeeze and excite attention module (Hu et al., 2018a) was inserted after the block output in the transition layer, which facilitates utilisation of multi-scale information.

For the full description of this method, please refer to Liao et al. (2023).

Unlike colonoscopy polyp segmentation, the DFUC 2022 segmentation challenge does not include real-time processing as a criterion. To obtain higher accuracy, a more complex decoder was selected, the Lawin decoder, to replace the original decoder of HarDNet-MSEG. The keypoint of the decoder of the Lawin Transformer is the proposed attention mechanism called Large Window Attention, which can capture multi-scale features and represent the segmentation result more precisely (see Fig. 1).

To consider the full dataset, 5-fold cross validation was used to obtain five sub-models, followed by test time augmentation to test different transformed images, using transformations such as vertical and horizontal flips. Finally, the average result values from the sub-models are used as the final output. The outputs are passed through the

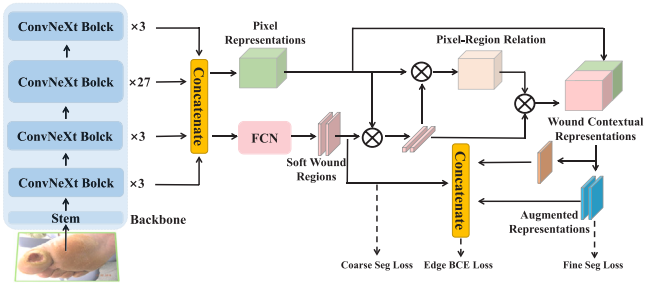

Fig. 2. Overview of Edge_OCRNet. On the basis of OCRNet, the backbone network is replaced by ConvNeXt, the outputs of the four layers of the backbone are merged and an edge loss is added.

Tanh function to generate a binary predicted mask and then rounded to {0, 1}. After completing these steps, we apply morphological operations to fill the holes within segmented regions to improve the true positive rate.

3.2. Yi et al. (2nd place, LkRobotAILab team)

The focus of Yi et al.'s approach (Yi et al., 2023) was to improve on the fine details of DFU segmentation predictions. First, a coarse-to-fine two-stage structure was used, similar to how the human visual system functions. Second, edge information was added from the DFU mask as additional supervisory information during training. For the coarse-to-fine structure, OCRNet (Yuan et al., 2019) was used as the baseline model. The first stage of the baseline is a simple FCN (Long et al., 2015). In this stage, the FCN is used to coarsely segment the DFU and the result is fed into the next stage as wound semantic information. In the second stage, the wound semantic information interacts with the pixel representation information to produce a more detailed segmentation result. To extract more robust pixel and semantic representational information, ConvNeXt (Liu et al., 2022), a state-of-the-art classification network was chosen as the backbone for the model. Additionally, the output features of the four layers of the ConvNeXt encoder were concatenated to enhance the model's perception of spatial information and to improve its generalisation of changes in the object scale. In the DFUC 2022 dataset, the diverse representation of DFU edges was noted. In order to further improve the DFU segmentation results, an "edge loss" loss function was added to constrain boundary information. The above improvements form the final model, namely Edge-OCRNet. Its structure is illustrated in Fig. 2. For a full description of this method, please refer to Yi et al. (2023).

3.3. Galdran et al. (3rd place, Agaldran team)

This approach was focused on a specific aspect of the foot ulcer segmentation problem: analysing the robustness to the potential absence of a DFU in the image (Galdran et al., 2023). In this case, robustness was understood as reliably handling images that might not contain any DFU, without creating false positives. Note that in DFUC 2022 (Kendrick et al., 2022), predicting a single DFU pixel on a DFU-free image would result in a DSC of 0. Therefore it becomes critical to avoid false positive detections in disease-free samples.

With the aim of training a model that reliably discards healthy images, we carried out a comprehensive analysis on the impact of a range of five popular segmentation loss functions, which were used to optimise the weights of an array of different architectures, all of which were double encoder–decoder networks, but with different architectural backbones (Galdran et al., 2021). Fig. 3 illustrates the architecture of the proposed method. As detailed in Section 5.3, the standard Cross-Entropy loss function was shown to be the most robust of all the loss functions tested with DFU-free images. Coupled with a five-fold

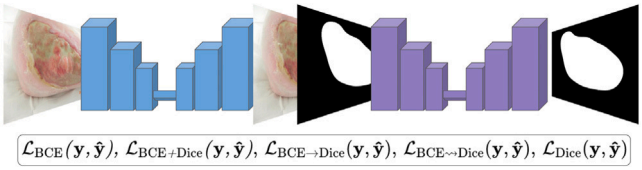

Fig. 3. Overview of the AGaldran Team approach. A double encoder–decoder network (Galdran et al., 2022) takes an image, generates a prediction and then uses the image with the prediction to refine the output. This architecture was optimised with five different loss functions in order to find out which option would work better in the absence of a DFU on the input image.

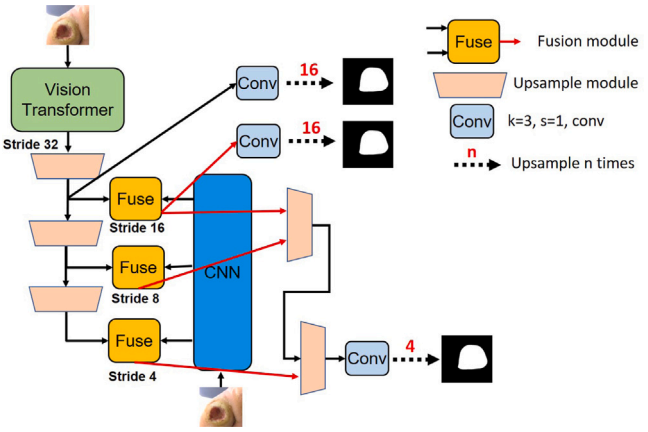

Fig. 4. Overview of the modified TransFuse model proposed by the ADAR-LAB team for DFUC 2022. The Transformer-branch encoder is CSwin-Base and the CNN-branch is ResNet-50. They are fused by Fusion modules at 3 different resolutions. Due to the additional upsampling module applied on the Transformer branch, feature maps can be decoded at a higher resolution compared to the original TransFuse model.

ensemble of a double FPN model with a ResNeXt101 backbone, this was the final submitted solution, as outlined in Section 4.3. The analysis shows that using the popular DSC loss for segmenting DFU would result in accurate delineations whenever an ulcer was present, but tended to generate spurious predictions when the image contained no DFU. This anomaly is likely due to the well-known miscalibration of models trained with Cross-Entropy loss (Mehrtash et al., 2020-12).

3.4. Chen et al. (4th place, ADAR-LAB team)

ADAR-LAB proposed a modified TransFuse model for DFUC 2022, which consists of a transformer branch, CNN branch, and fusion modules. The overview of their proposed architecture is shown in Fig. 4.

The CSWin-Transformer was selected as the backbone of the Transformer branch (Dong et al., 2022) due to its state-of-the-art performance with publicly available pre-trained weights. An additional upsampling module was applied in the Transformer branch to allow for the feature maps to be decoded at a higher resolution, so that the error on the edges can be reduced.

For the CNN branch, ResNet-50 and HarDNet-68 (He et al., 2016; Chao et al., 2019a) were considered for the backbone to extract local features. Limiting the model size is helpful to decrease the memory overhead during training. Through our experiments, we found that ResNet-50 demonstrated higher performance in validation, so it was adopted as the CNN-branch backbone in the proposed model architecture.

The function of the Fusion module is to apply attention to and combine the outputs of the Transformer and CNN branches. Two kinds of attention modules are applied: the squeeze-and-excite (SE) block, a channel-attention technique, and the convolutional block attention module (CBAM), a spatial-attention technique (Hu et al., 2018b; Woo

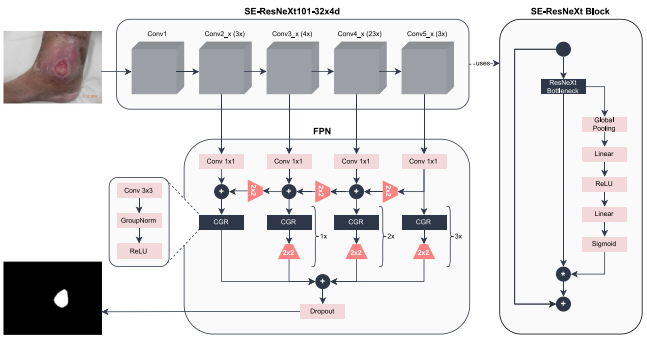

Fig. 5. Overview of the Feature Pyramid Network (Lin et al., 2017) architecture, utilising a ResNeXt (Xie et al., 2017) backbone with a Squeeze-and-Excitation (Hu et al., 2020) module, partially adapted from illustrations from the original works.

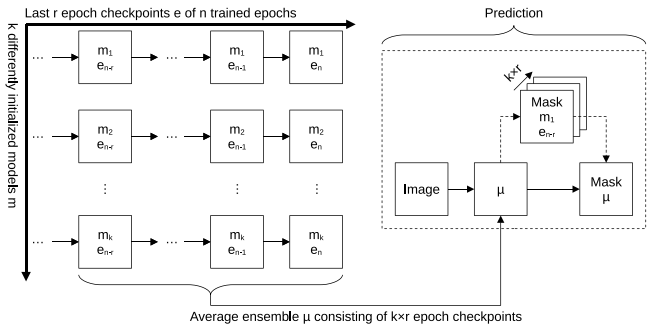

Fig. 6. Illustration of the ensemble approach used by the FHDO team: The last r checkpoints of n training epochs of k differently initialised models are used in an ensemble μ for averaged predictions.

et al., 2018). Because the feature maps from the CNN branch contain local information but lack global features, the CBAM block is more suitable. In contrast, since feature maps from the Transformer branch contain global features, but less local features, the SE block is considered to be a better fit. The design of our proposed method is in contrast to the design of the fusion module in the original TransFuse model, where the SE block is used for the CNN branch and the CBAM block is used in the Transformer branch. Full details of the proposed method are described in Chen et al. (2023).

3.5. Brüngel et al. (5th place, FHDO team)

Prior work during the DFUC 2021 (Bloch et al., 2022) has proven the potential of Generative Adversarial Network (GAN) (Goodfellow et al., 2014)-generated synthetic images for dataset enrichment. In such cases, DFU infection and ischaemia classification performance was demonstrably improved. The approach used by team FHDO during the DFUC 2022 (Brüngel et al., 2023) again relied on such a strategy to investigate the effects of synthetically generated DFU images on DFU segmentation performance. However, for this new segmentation task, the implementation differs in accordance to the nature of the segmentation problem.

Usually, conditional GANs (Mirza and Osindero, 2014) should be preferred for the task of segmentation dataset enhancement. Masks used for synthetic image generation can directly serve as ground truth labels. Furthermore, masks can be shaped arbitrarily, enabling measures for increasing robustness against non-standard shape representations. However, preliminary experiments using the DFUC 2022 dataset with conditional GAN implementations did not yield adequate synthetic image quality. This could mainly be ascribed to the limited number of training images ($n = 2000$) with a predominant amount of very small DFU wound instances, as low data-efficiency of current conditional GAN implementations is a bottleneck. The lack of spatial information/complexity in single-class DFU segmentation problems also hinders conditional GANs in achieving high-quality generation results. Non-DFU tissue and other human body parts are equated with overall background features, making differentiated feature learning of such areas highly challenging. To address such challenges, the proposed method used StyleGAN2+ADA (Karras et al., 2020), a member of the unconditional StyleGAN (Karras et al., 2019) family that applies Adaptive Discriminator Augmentation (ADA) to achieve a high data-efficiency.

The proposed DFU segmentation approach involved models with a Feature Pyramid Network (FPN) (Lin et al., 2017) architecture, using an SE-ResNeXt101-32x4d backbone as variant of ResNeXt (Xie et al., 2017) including a Squeeze-and-Excitation (SE) (Hu et al., 2020) module. An overview of the proposed architecture is shown in Fig. 5.

Predictions were inferred using a large ensemble of different model checkpoints, as shown in the schematic in Fig. 6. A total of $k = 5$ differently initialised models were trained using a five-fold cross validation approach. The last $r = 5$ epoch checkpoints of n epochs were persisted and utilised in an ensemble μ, consisting of $k \times r = 25$ checkpoints. The $k \times r = 25$ predictions of these for an image were averaged to a final mask, oriented towards Polyak–Ruppert averaging (Polyak, 1990; Ruppert, 1988).

Essential implementation details are summarised in Section 4.5, with more detailed elaborations given in Brüngel et al. (2023).

4. Implementation

4.1. Yllab team implementation

During the training stage, ImageNet pre-trained weights were used to initialise the backbone, followed by training for 300 epochs using the AdamW optimiser. The batch size was set to 6, with a learning rate of $l = 1e{-}4$ with a cosine-annealing scheduler. To preserve the original image aspect ratio, the training images are zero-padded into square dimensions then randomly resized between 384 × 384 and 640 × 640 pixels. To allow for the model to learn more features across the different examples, data augmentation methods were employed during training including random vertical flipping, horizontal flipping, and cropping, etc.

During the testing phase, Test Time Augmentation was with the five sub-models respectively, with the predictions averaged to form the final results.

4.2. LkRobotAILab team implementation

The solution relies on MMSegmentation,[1] an open source semantic segmentation toolbox based on PyTorch, which is discussed in Section 3.2. All models were trained and tested on a single NVIDIA GeForce RTX™ 3090 24G. The following is a discussion of the specific implementation details.

In the training phase, pre-trained ImageNet (Deng et al., 2009) classification networks were used as backbones. The best performing of these pre-trained models was ConvNeXt-XL (Liu et al., 2022) pre-trained on ImageNet-21K. The AdamW optimiser was used with a learning rate to 8e-5, weight decay of 0.05, a warmup step rate of 1500, a warmup ratio of 1e-6, a batch size of 4, and training iterations set to 60K if not specified. For training, the original DFUC 2022 training images were used as input (480 × 640 pixels) together with multi-scale examples used for data augmentation with an image size

[1] MMSegmentation: https://github.com/open-mmlab/mmsegmentation/ (accessed 2023-02-03).

distribution interval of (0.5, 2.0) in steps of 0.25. Horizontal flipping (with 0.5 probability), random cropping (cropping size 512 × 512, maximum crop rate 0.75), and Photometric Distortion augmentation methods were also used to enhance the training data and to improve the generalisation ability of the model.

The use of multi-scaling and random cropping strategies during the training phase allowed the model to benefit from the FixRes effect (Touvron et al., 2019) whereby performance is improved by using larger images during testing. During experiments, we observed that size range $S_{tr}/S_{infer} = [1.2, 1.5]$ can produce better results in the DFUC 2022 segmentation task when using this training strategy. An input image size of 576 × 768 pixels was selected for the prediction. To alleviate the difference between the training domain and the test domain, test images were processed with gamma correction to increase the brightness of the test set. The ratio $r_{gamma} = RGB^{train}_{mean}/RGB^{test}_{mean}$ was used to determine gamma values in test images. In addition, multi-scale testing and TTA (horizontal and vertical flips) were also used during the inference of test images.

The DFUC 2022 segmentation task is a binary classification task for each pixel. At the time of prediction, the pixel is considered to belong to the DFU if its prediction score is greater than a threshold (t=0.655), otherwise, it is considered to be the background.

4.3. AGaldran team implementation

This approach adopted a Feature-Pyramid Network architecture as part of a double encoder–decoder variant, as in the authors prior works (Galdran et al., 2021). Several backbone encoders, optimised with five different loss functions were also used. Each model was pre-trained using Imagenet weights, which were then optimised using the Adam optimiser with a learning rate of $l = 3e-4$, a batch-size of 4, and an image size of 640 × 512. During training, images were augmented using common image processing operations (random rotations, translations, scalings, vertical/horizontal flipping, and contrast/saturation/brightness changes.). During the testing phase, Test-Time Augmentation was utilised.

The above process resulted in selecting a ResNext101 backbone trained and optimised with the Cross-Entropy loss function, which resulted in the highest performance, both in an internal five-fold cross-validation setup, and also when submitting to the public DFUC 2022 validation leaderboard. The segmentation results from the final test set, using the resulting five-fold model ensemble, were submitted to the DFUC 2022 test leaderboard.

4.4. ADAR-LAB team

Prior to training the segmentation network, an RPN with a ResNet-50 backbone was trained for 10,000 iterations (Girshick, 2015). The AdamW optimisation algorithm was employed to optimise parameters with a learning rate of 10^{-5}. The batch size of each iteration was set to 32, sampled randomly from 1800 images in the training dataset, with the remaining 200 images used as validation data.

For the segmentation task, two stages were employed in the training process. First, we initialised of two backbones (He et al., 2016; Dong et al., 2022) using ImageNet pre-trained weights. The models were then trained with 1800 images using the AdamW optimiser, with a learning rate of 3×10^{-5}, and a batch size of 8 for 100 epochs. The loss function is the same as defined in the PraNet implementation (Fan et al., 2020). The model with the highest validation metrics was saved, and then used as the pre-trained weights for the next phase.

Using the best-performing parameters from the first stage of training, the second stage of training was completed using all 2000 training images for 50 epochs, or 20 epochs when the multi-scaling method was applied. In this stage, the batch size was set to 12, in which 4 comprised of resized inputs, and another 4 were cropped from one of corners, with the remainder cropped according to the RPN results. Two different optimisers were used: (1) SGD (used only for the CNN-branch backbone in the TransFuse model) with a momentum of 0.9, a learning rate of 10^{-4}, scheduled by a cosine-annealing scheduler; and (2) AdamW with a learning rate of 3×10^{-5}, also utilising a cosine annealing scheduler and employed for the rest of parameters in the model. The learning rates of both optimisers would decay to 0.1 of the initial learning rate at the end of training.

4.5. FHDO team implementation

The approach described in Section 3.5 utilised the Segmentation Models PyTorch[2] (SMP) as a wrapper framework for PyTorch (Paszke et al., 2019)-based segmentation implementations, and StyleGAN2+ ADA (Karras et al., 2020) for synthetic DFU image generation. The following subsections summarise implementation aspects, for which further details are reported in Brüngel et al. (2023).

4.5.1. Ensembles of base models and extended models

As intensities of DFU instance border regions in the training set ground truth masks ranged from 0 to 255, masks were binarized using a threshold of ≥ 128. A total of 39 duplicate image pairs were identified with slightly differing ground truth were merged, keeping one instance with logically `OR`-ed corresponding masks. The resulting training dataset comprised of 1961 images with 2262 DFU instances.

For all baseline and extended segmentation models, nearly identical training configurations were used. An FPN (Lin et al., 2017) architecture with SE-ResNeXt101-32x4d (Hu et al., 2020; Xie et al., 2017) backbone was used, pre-trained on ImageNet (Deng et al., 2009), using the *sigmoid* activation function. Adam (Kingma and Ba, 2015) was chosen as the optimiser with an initial learning rate of 0.0001, with DSC as the loss function. All models were trained on a single NVIDIA®V100 16 GB,[3] using a batch size of 24. The only differences between baseline and extended models were the number of trained epochs, the learning rate schedule, and, most decisively, the training set variant (baseline or extended) they were trained on. Baseline models were trained for 150 epochs, dropping the learning rate to 0.00005 at epoch 100 and to 0.00001 at epoch 135. Extended models were trained for 150 epochs, dropping the learning rate to 0.00001 at epoch 120.

A consistent augmentation pipeline was applied for training, implemented using the Albumentations library (Buslaev et al., 2020). Augmentations and parameters were chosen to not distort DFU representations beyond domain-typical variance, thus excluded methods such as colour shifts or channel drops. If not stated otherwise, the default parameterisation of operations was used. The pipeline first applied guaranteed random cropping (`RandomCrop` with `width/height` = 352, p = 1.0), followed by random image flipping (`Flip`, $p = 0.5$) as well as shifting, scaling, and rotating (`ShiftScaleRotate`, $p = 0.5$). To distort images, either grid distortion, elastic transformation (`GridDistort`, or `ElasticTransform`, $p = 0.5$) were applied randomly. Brightness and contrast were also modified, applying either contrast-limited adaptive histogram equalization (CLAHE), random gamma, or random brightness and contrast (`CLAHE`, `RandomGamma` or `RandomBrightnessContrast`, $p = 0.5$). Random sharpening or (motion) blurring (`Sharpen` or `Blur` with `blur_limit` = 8 or `MotionBlur` with `blur_limit` = 8, $p = 0.5$) was also applied. Gaussian noise (`GaussianNoise`, $p = 0.5$) was added as final step.

The baseline models were trained on the clean baseline training set variant with 1961 images. These were then used to pseudo-label synthetic images, generated as described in Section 4.5.2. Extended models were then trained on the synthetically enriched training set with 5961 images (+4000 pseudo-labelled synthethic images).

[2] Segmentation Models PyTorch library: https://github.com/qubvel/segmentation_models.pytorch/ (access 2023-01-29).

[3] https://www.nvidia.com/en-us/data-centre/v100/ (2023-01-29).

Table 1
A comparison of the methods proposed by the winners. Note that all methods adopted data augmentation and pre-trained models in their implementation. We observe that the backbone selections are varied for the participants. *TTA: Test Time Augmentation*.

Team	Augmentation	Method	Backbone	Post-processing
Yllab	Yes	HarDNet-DFUS	HarDBlkV2 & Lawin Transformer	Average of TTA
LkRobotAILab	Yes	Edge-OCRNet	ConvNeXt	Multi-scale and TTA
AGaldran	Yes	Double Encoder-Decoder	ResNeXt101	5-fold models ensemble
ADAR-LAB	Yes	Modified Transfuse model	CSWin-Transformer & ResNet50	Voting of TTA
FHDO	Yes	Feature Pyramid Network	SE-ResNeXt101-32 × 4d	Ensemble approach

Table 2
Results of five submissions using HarDNet-DFUS in the final testing phase of DFUC 2022.

Model	Dice
HarDNet-DFUS+Deep1+Boundary	0.7237
HarDNet-DFUS+Deep1+Boundary (w/ hflip)	0.7243
HarDNet-DFUS+Deep1+Deep2+Boundary (w/ hflip)	0.7273
HarDNet-DFUS+Deep1+Deep2+Boundary (w/ vflip)	0.7275
HarDNet-DFUS+Deep1+Deep2+Boundary (w/ vhflip)	**0.7287**

Table 3
The performance of LkRobotAILab's model in the final testing phase of DFUC 2022.

Model	Dice	Jaccard
OCRNet+HRNet-48	0.7057	0.6028
OCRNet+ConvNeXt-XL	0.7219	0.6194
OCRNet+ConvNeXt-XL+Edge-loss	0.7226	0.6207
OCRNet+ConvNeXt-XL+Edge-loss+TA[a]	**0.7280**	**0.6276**

[a] Includes TTA and multi-scale testing.

Table 4
Performance of different models trained by team AGaldran on a variety of loss functions on the DFUC 2022 testing set.

Model + Loss Function	DSC
TTA 1: No TTA	72.33
TTA 2: Rotation 15°	72.40
TTA 3: Colour Jittering	72.56
TTA 4: Horizontal Flip	72.61
TTA 5: Horizontal Flip+Colour Jittering	**72.63**

All segmentation model predictions were inferred at a confidence threshold of 50 % and had the same weight in averaged ensembles. Simple post-processing was applied for all baseline ensemble predictions on the validation set, and pseudo-labels for synthetic images, involving instance filtering by size. Instances detected by a contour finding algorithm (Suzuki and Abe, 1985) were removed when measuring < 1 ‰ of the whole image area. This was only applied to predictions with more than one instance. Finally, opening was applied with a 2×2 square kernel for size filtering artefact removal. Whether this procedure was applied for a submission or not is stated in the reported results in Section 5.5.

4.5.2. Synthetic image generation and pseudo-labelling

A StyleGAN2+ADA (Karras et al., 2020) generation model was trained for 1000 steps on 512×512 px centre crops of the pre-processed training set with activated mirroring amplification. This involved four NVIDIA®V100 16 GB GPUs, enabling a batch size of 32. The default 512 px configuration with Flickr Faces HQ (Kazemi and Sullivan, 2014) pre-trained weights as well as the default ADA settings were used. A minimum Frechet Inception Distance (FID) of 19.09 was achieved at 880 steps, using respective weights generation of 4000 synthetic images using the seeds $0 - 3999$.

Synthetic images were then pseudo-labelled using the previously created baseline ensemble, applying the post-processing procedure described in Section 4.5.1. Synthetic images yielded as PNG were then converted to JPEG with the same compression level as the DFUC 2022 testing set images. These, together with their pseudo-labels, were then added to the training dataset.

Table 1 compares the proposed methods and the implementations. All the methods deployed data augmentation and pre-trained models. Various post-processing techniques were used to produce the final results, with 3 teams using Test Time Augmentation (TTA) and two using ensemble approaches.

5. Results

5.1. Yllab results

Table 2 shows the results of five team Yllab submissions during the final testing phase. We experiment with different deep supervision and TTA methods to improve the performance of our model. There are two deep supervision losses, called Deep1 and Deep2, and a boundary loss, called Boundary. With the addition of deep supervision losses, boundary loss, and the horizontal flip with a vertical flip TTA method, HarDNet-DFUS achieved 0.7287 mean DSC and ranked first among all teams.

5.2. LkRobotAILab results

In Table 3, the ablation study of the improvement of our method is shown. By using a robust backbone, adding edge loss, and using several TTA methods, our solution achieved 0.7280 (rank 2) mean DSC and 0.6276 (rank 1) mean Jaccard during the testing phase.

5.3. AGaldran results

The analysis of this team was guided towards understanding which loss function, within a set of five candidates, was more capable of dealing with disease-free images. The considered candidates were the standard Cross-Entropy (CE, $\mathcal{L}_{\text{BCE}}$) and DSC loss ($\mathcal{L}_{\text{Dice}}$), together with a series of three different combinations, namely: adding them ($\mathcal{L}_{\text{BCE+Dice}}$), linearly interpolating from CE to DSC during training ($\mathcal{L}_{\text{BCE}\rightarrow\text{Dice}}$), and training for 90% of the epochs with $\mathcal{L}_{\text{BCE}}$and then switching to DSC ($\mathcal{L}_{\text{BCE}\nrightarrow\text{Dice}}$).

Internal cross-validation results indicated that $\mathcal{L}_{\text{BCE+Dice}}$ and $\mathcal{L}_{\text{BCE}\nrightarrow\text{Dice}}$ could achieve high performance when evaluated on images that would always contain ulcers, which composed the original training set. However, by assessing performance upon submission to the public validation set (which contained lesion-free samples) during model development, we found that this trend was reverted and the CE loss $\mathcal{L}_{\text{BCE}}$resulted in highest performance, whereas models trained minimising $\mathcal{L}_{\text{BCE+Dice}}$ and $\mathcal{L}_{\text{BCE}\nrightarrow\text{Dice}}$, as well as $\mathcal{L}_{\text{Dice}}$, resulted in a drastic degradation in DSC. Once the $\mathcal{L}_{\text{BCE}}$ loss was adopted as the best solution, a ResNeXt-101 encoder coupled with a Feature-Pyramid-Network was trained and corresponding segmentation results were submitted using the hidden test set. Several post-processing steps were tested to ascertain which was the most appropriate TTA scheme, considering several combinations: (1) no TTA, (2) only flip image horizontally, (3) only Rotate the image 15°, (4) only colour jittering, (5) flip image and colour jittering. The results shown in Table 4 indicate that no extreme differences in performance occurred, although the last combination was +0.30 DSC which indicates superior performance compared to not using TTA at all, which can be considered as a relevant difference since the winner of the challenge was +0.24 DSC over our approach.

Table 5
The results of submissions for team ADAR-LAB after the final testing phase.

Model	DSC
None	0.7270
focal loss	0.7280
focal loss + multi-scale	**0.7285**

Table 6
Team FHDO results for baseline and extended model ensembles, best results per challenge phase are highlighted.

Ensemble	Post-proc.	Dice	Jaccard	FNE	FPE
Submissions for validation set					
Baseline	Yes	0.6895	0.5880	0.2693	0.2493
Extended	Yes	**0.6971**	**0.5974**	**0.2578**	**0.2466**
Final submissions for testing set					
Extended	No	**0.7169**	**0.6130**	**0.2453**	**0.2145**
Extended	Yes	0.7136	0.6086	0.2470	0.2195

5.4. ADAR-LAB results

In the ablation study, we obtained experiment results of the methods we applied in this challenge with fine-tuned parameters after the final testing phase, which resulted in a DSC of 0.7270 on the testing dataset. Then, we gradually introduced a number of common methods, and the results are shown in Table 5. After adding focal loss to our loss function, there was an improvement of 0.001 in DSC. Training the model with randomly resized images further improved DSC by 0.0005.

5.5. FHDO team results

Synthetic DFU representations generated by the unconditional StyleGAN2+ADA model showed an overall good quality, in terms of realism, and a broad variety. Examples are shown in the left column of Fig. 7. Representations of feet mostly ranged from anatomically unobtrusive and presumably healthy feet with absence of DFUs, up to malformed or partially amputated feet with highly diverse DFU representations. Anatomically implausible representations of feet were also generated, e.g., highly elongated body parts or multiple extra-toes. However, such examples constituted the minority of cases. DFU representations ranged from small and early stage wounds, over medium-sized and well demarcated areas, up to large-sized areas with complex tissue composition. All three main tissue types, granulation, slough, and necrosis, were present either uniformly, or as part of tissue combinations. Wound bed depths ranged from deep hole-like, over shallow, up to protruded hypergranulation-like manifestations. Common accompanying symptoms such as rhagades, macerated wound borders, hyperkeratotic or flaking skin layers, and reddened or discoloured peri-wound areas were generated as well, showing high levels of detail. Backgrounds typically showed white cloth, blue foil, or mixes of both which are characteristic for the DFUC 2022 dataset. Pseudo-ground truth for synthetic images created by the baseline model ensemble showed good levels of consistency. Examples of binary masks are shown in the middle column of Fig. 7 with corresponding image cutouts in the right column. Further details and numerous samples of synthetic images with created pseudo-ground truth are reported in Brüngel et al. (2023).

Performance of the baseline and extended model ensembles are reported in Table 6. The upper part reports results on the validation set, the lower part results on the testing set. During the validation phase, both the baseline and the extended ensemble were evaluated, with both using the described post-processing procedure. In this phase, the extended model ensemble performed consistently better than the baseline model ensemble, achieving notably higher DSC and Jaccard values as well as notably lower FNE and FPE. During the testing phase, only the extended model ensemble was used for submissions,

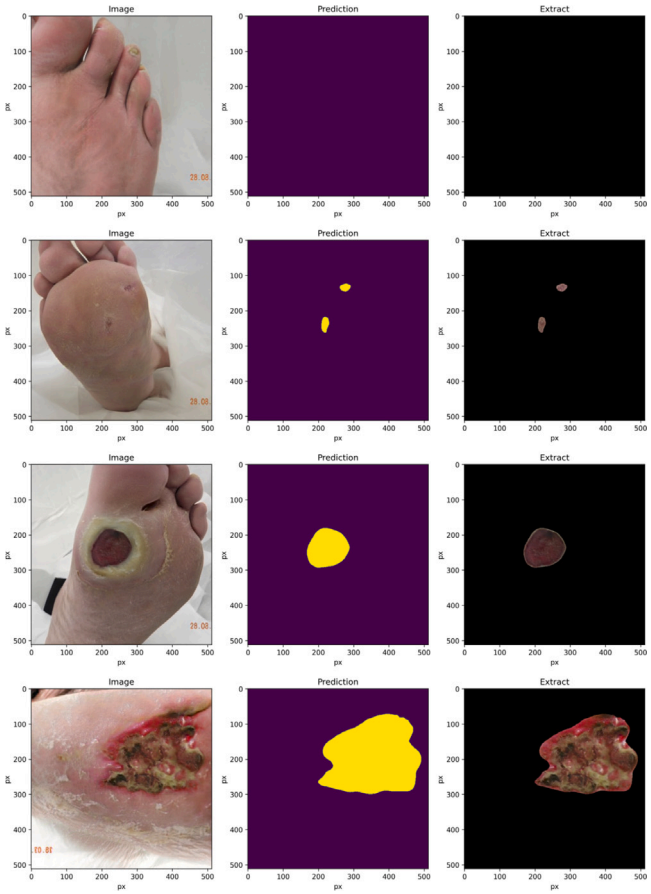

Fig. 7. Examples of generated synthetic images with pseudo-labels predicted by the baseline ensemble, as generated by team FHDO: The first row shows a presumably healthy foot, the second row shows multiple small-sized and shallow DFU instances, the third row shows a medium-sized and well-demarcated DFU with pronounced maceration in the periwound, and the fourth row shows a large-sized DFU instance with mixed tissue and non-uniform edges.

with and without involvement of the post-processing procedure. The final results show that the extended model submission without post-processing performed best, achieving a DSC of 0.7169, a Jaccard value of 0.6130, an FNE of 0.2453, and an FPE of 0.2145. Further results of post-challenge evaluations are reported in Brüngel et al. (2023), with particular attention to potential overfitting in the final results.

5.6. Comparison of the best model from each method

On the final submission leaderboard, the performance of the methods from the DFUC 2022 winners with DSC values > 0.70 on non-DFU, small, medium and large DFU regions are illustrated on Fig. 8. The top-10 results is summarised in Table 7. Noted that group "seoyoung" did not submit a paper to describe their method, therefore, was not considered in the analysis. The fifth place was awarded to FDHO Team. The number of test set submissions for the top-5 teams was as follows: Yllab ($n = 5$), LkRobotAILab ($n = 4$), AGaldran ($n = 5$), ADAR-Lab ($n = 5$), FHDO ($n = 5$). The mean DSC and Jaccard values for the top-5 teams are 0.7261 and 0.6251 respectively, with a standard deviation of 0.0026 and 0.0027 respectively.

6. Comprehensive analysis

In this section, we demonstrate the ability of the networks by performing a comprehensive analysis of the segmentation results. First, we analyse the effect of ensemble methods in DFU segmentation; Second,

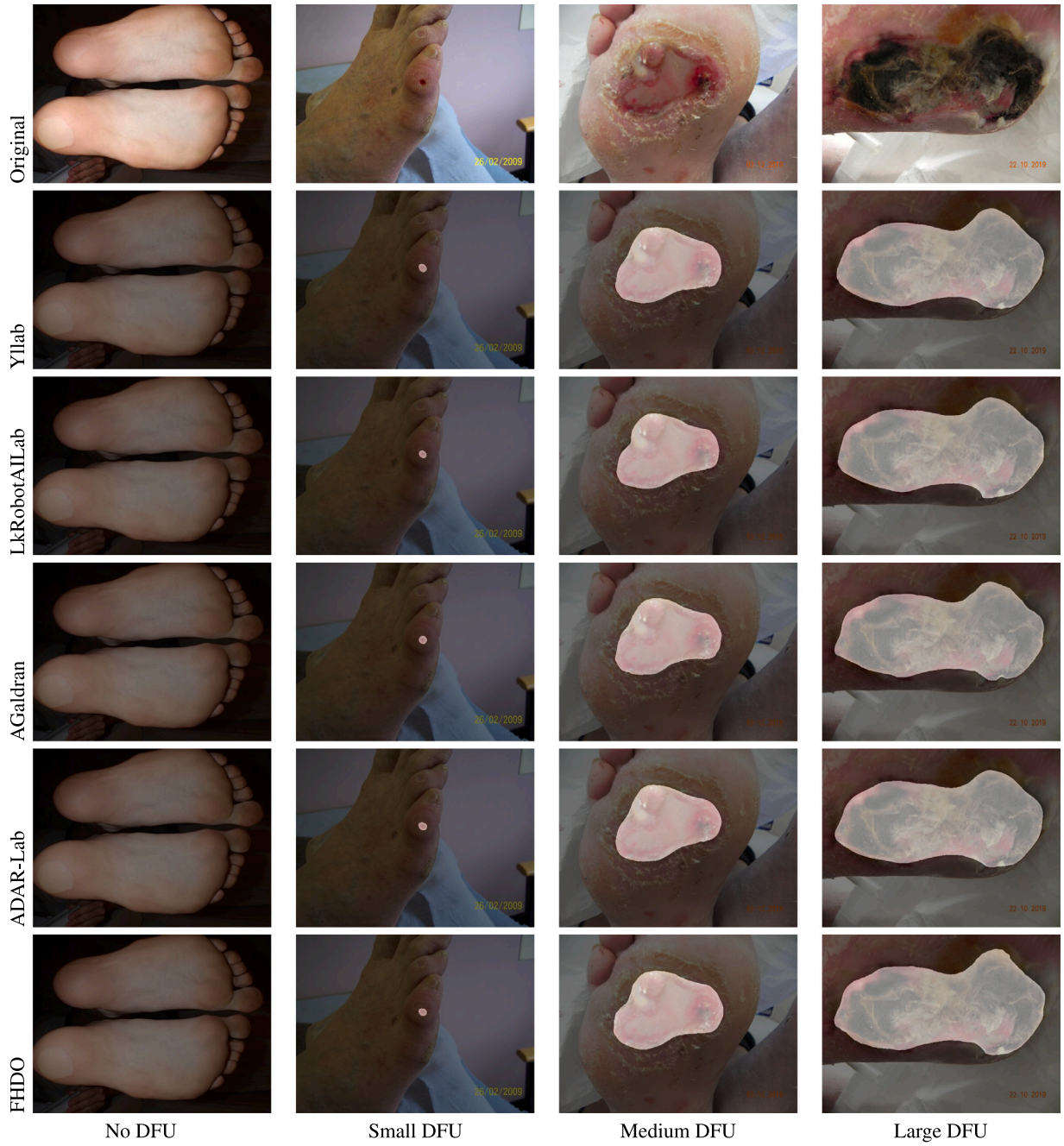

Fig. 8. Comparison of a selection of prediction results from participating teams overlayed on the test images.

we perform statistical analysis on the winning teams' results; Third, we implement region-based measurement; and finally, we investigate the relationship between DSC values and region sizes.

6.1. Ensemble methods

A common approach to improve segmentation metrics is to ensemble a series of the best performing models, with each model capable of identifying alternative features. For instance, some models may exhibit superior performance in the identification of infected regions, while other models may show better performance in segmenting early stage DFU. Additionally, using an ensemble of models can help to avoid false detections by allowing the networks to vote on active regions which can create a more robust system. In this section, we analyse the ability of the networks to cohesively segment DFU regions.

We perform an analysis of three different ensemble types for segmentation, namely:

- Union: if one method indicates that the pixel is DFU then the ensemble will classify as DFU
- Voting: if half or more of the methods indicate that the pixel is DFU then the ensemble will classify as DFU
- Intersection: if all methods voted the pixel as DFU, then the ensemble will classify it as DFU

For the ensemble analysis, we take the best-performing networks from the DFUC 2022 winners with DSC values > 0.70, and post-process their results with the ensemble methods. We received the binary mask predictions from Yllab (Liao et al., 2023), LkRobotAILab (Yi et al., 2023), AGaldran (Galdran et al., 2023), ADAR-LAB (Chen et al., 2023)

Table 7
The top-10 participating teams for DFUC 2022, starting with the best DSC value.† = higher score is better; ⊎ = lower score is better. **bold** indicates the best overall result.

Team	Metrics			
	Dice †	Jaccard †	FPE ⊎	FNE ⊎
Yllab	**0.7287**	0.6252	0.2048	0.2341
LkRobotAILab	0.7280	**0.6276**	0.2154	0.2261
AGaldran	0.7263	0.6273	0.2262	**0.2210**
Adar-Lab	0.7254	0.6245	**0.1847**	0.2582
seoyoung	0.7220	0.6208	0.1925	0.2584
FHDO	0.7169	0.6130	0.214	0.2453
GP_2022	0.6986	0.5921	0.2065	0.2778
DGUT-XP	0.6984	0.5945	0.2523	0.2379
IISlab	0.6974	0.5926	0.2163	0.2734
AGH_MVG	0.6725	0.5690	0.2555	0.2830

Fig. 9. The results of ensemble of the winning team's images with different cases, Intersection (left), Union (middle) and Vote (Right).

Table 8
Results of each of the ensemble methods. *Italics* indicates the best performing method for each ensemble method, and **bold** indicates the best overall metric value.

Ensemble Method	Metrics			
	DSC	Jaccard	FPE	FNE
Intersect1	*0.7264*	*0.6263*	*0.2440*	0.2025
Intersect2	0.7202	0.6209	0.2673	0.1907
Intersect3	0.7179	0.6191	0.2757	0.1874
Intersect4	0.7122	0.6134	0.2866	***0.1848***
Union1	*0.7302*	*0.6270*	0.1763	*0.2556*
Union2	0.7279	0.6248	0.1654	0.2677
Union3	0.7263	0.6231	0.1544	0.2771
Union4	0.7196	0.6146	***0.1477***	0.2925
Vote1	0.7264	0.6263	0.2440	*0.2025*
Vote2	0.7318	0.6320	0.2133	0.2224
Vote3	0.7299	0.6307	0.2285	0.2128
Vote4	***0.7322***	***0.6324***	*0.2099*	0.2253

and FHDO (Brüngel et al., 2023). We then process the method with the 3 techniques and visualise the results (see Fig. 9).

Table 8 illustrates how each method optimises separate metrics. In the case of intersection, we see a significant reduction of FNE with the highest score equal to the lowest on the next ensemble method, highlighting how the removal of none intersecting pixels highlights core DFU related features. Whereas, in the case of FPE, Union reduces the FPE for all ensemble methods below that of other ensemble methods, showing that the different methods highlight different sections and features of the DFU regions. Finally, the voting system shows the best performance for DSC and Jaccard, highlighting the ability of the combined approach to improve prediction overlap. Additionally, we observe that the values improve with each additional method, highlighting that the diversity in the combined network predictions allow for significantly improved segmentation. In-contrast, the opposite correlation is also demonstrated — if the ensemble improves FPE then a reduction in FNE occurs.

While ensembling the predictions of the top teams appears to demonstrate marginal improvements to the results, we perform further analysis on different test sets. Fig. 10 illustrates the mean DSC of winning teams and ensemble methods on 3 test sets based on the ratio of the DFU area to the image area. We split the test set based on the ratio values in the range of (0,0.01), (0.01,0.05), and (0.05,0.10). We observe that the algorithms are less accurate in segmenting small DFU regions, but gradually improve as DFU areas increase. The union ensemble method demonstrated the worst performance for small DFU areas. Overall, this analysis demonstrates that the most difficult cases for DFU segmentation are images with small DFU regions, in particular for those with a ratio < 0.01.

Table 9
Comparison of segmentation results using region-based measurement with a threshold of $DSC > 0.5$. **bold** indicates the best overall result.

Methods	Metrics			
	Accuracy	Recall	Precision	F1-Score
Yllab	0.6222	0.7890	0.7445	0.7661
LkRobotAILab	0.6473	0.7925	0.7770	0.7847
AGaldran	**0.6706**	0.7783	**0.8265**	**0.8017**
Adar-Lab	0.6566	0.7796	0.8040	0.7916
FDHO	0.6108	**0.7961**	0.7218	0.7572

6.2. Statistical analysis

We conduct an analysis based on DSC values, but not on IoU as it is correlated with DSC. First, we determine if the multivariate sample means are equal by performing a MANOVA. MANOVA was selected as it improves on the capabilities of analysis of variance (ANOVA) by using multiple dependent variables simultaneously.

Fig. 11 illustrates the boxplots determined using test set DSC values for the best performing teams. Based on ANOVA we found that there were no significant differences between the mean DSC values obtained for each method (p-value =0.42).

We observed a strong linear correlation between the test DSC values obtained by the best performing teams, with pairwise Pearson correlation coefficients equal to approximately 0.85 (p-values<0.001). In particular, Fig. 12 illustrates the relationship between the DSC values obtained for the Yllab and LkRobotAILab teams. Although these two methods achieved similar DSC values, Fig. 12 shows that the networks misdetected different DFU images (DSC values of 0). This finding motivates the utilisation of ensemble methods in the previous sections since there were cases correctly segmented by one of the methods, but missed by the other.

6.3. Region-based analysis

The metrics used for DFUC 2022 are image-based, rather than region-based, i.e. the accuracy of each segment of the ulcer. However, in medical practice, the measure of true positives and false positives, based on each region, are used to measure the reliability of any computer aided tools. Therefore, we complete an analysis on region-based analysis. A region is deemed to be a true positive if the DSC value of the ground truth region and the predicted region is above a certain threshold. We compare the performance by calculating the following performance metrics: accuracy, recall, precision and F1-score.

The challenge methods segmented different numbers of regions, with Yllab predicting 2380 regions, LkRobotAILab predicting 2291 regions, AGaldran predicting 2115 regions, Adar-Lab predicting 2178 regions, and FHDO predicting 2477 regions. Fig. 13 shows the relationship between the metrics and the DSC threshold values. In addition, Table 9 compares the obtained results using region-based measurement for the threshold of $DSC > 0.5$, which corresponds to the case where at least half of the predicted region overlaps with the ground truth segmentation mask. It is noted that AGaldran team achieved higher results compared to the other methods, especially with respect to precision.

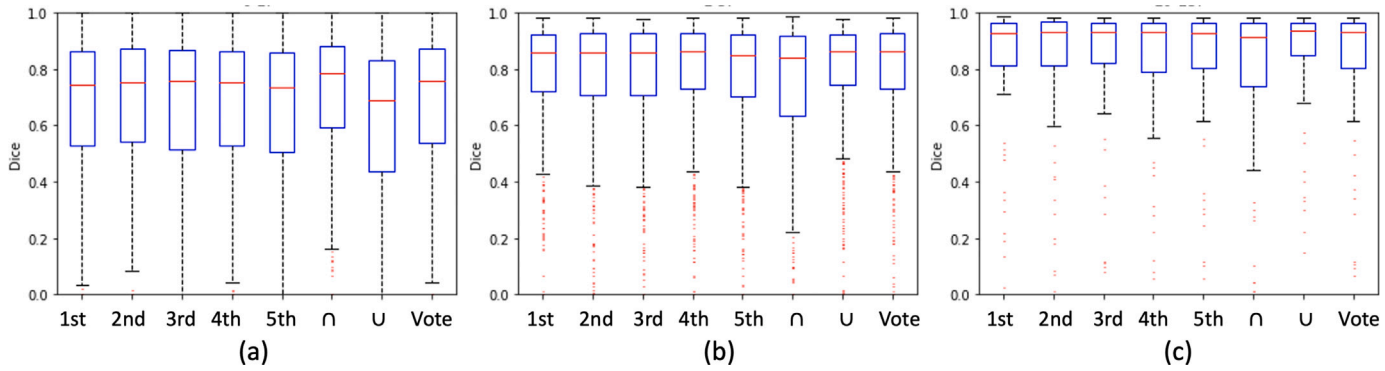

Fig. 10. Boxplots comparing the DSC values of the winning teams and ensemble methods on 3 test sets with different ratios of DFU area to image area, with (a) ratio in the range of (0,0.01); (b) ratio in the range of (0.01,0.05); and (c) ratio in the range of (0.05,0.10).
Note: 1st–5th indicate the 5 winning teams, ∩, ∪ and Vote indicates ensemble methods of intersection, union and vote, respectively.

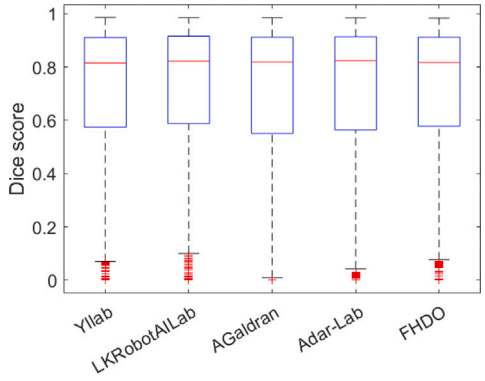

Fig. 11. Boxplots presenting the test set DSC values achieved by the higher performing teams in DFUC 2022.

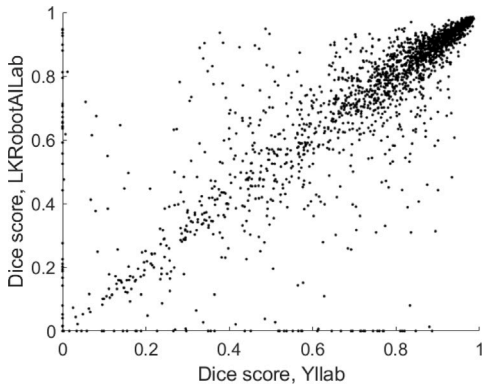

Fig. 12. Graph presenting the linear relationship between the DSC values obtained for the two higher performing teams in DFUC 2022 (Yllab and LkRobotAILab) with a Pearson linear correlation coefficient equal to 0.87.

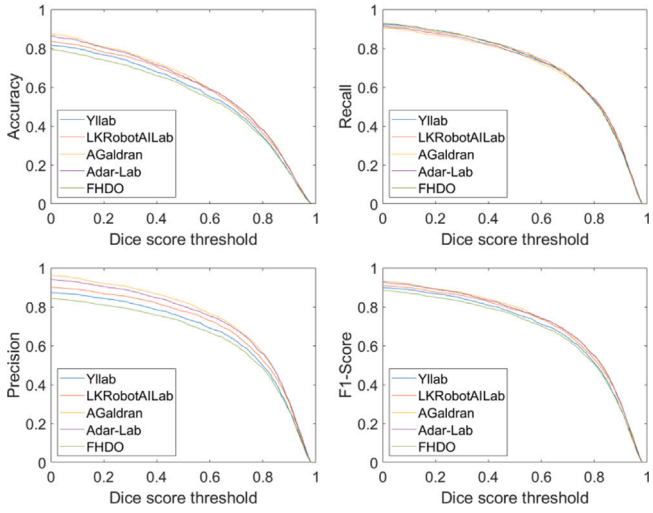

Fig. 13. Results of the region based analysis for different DSC threshold values.

6.4. Region size analysis

Previous studies reported that the region based loss functions (e.g. those utilising DSC) induce a bias towards a specific region size (Maier-Hein et al., 2020b). To assess this phenomenon with respect to the DFUC 2022 winner test set results, we investigated the relationship between DSC values and region sizes. First, we performed correlation analysis and found positive and significant (p-values<0.001) correlation between the DSC values and region sizes, with Spearman's rank correlation coefficients equal to 0.43, 0.40, 0.42, 0.43 and 0.40 for Yllab, LKRobotAILab, AGaldran, Adar-Lab and FHDO team, respectively. Fig. 14a) illustrates the relationship between the test DSC values and region sizes obtained for the network from the Yllab team. For visualisation, region sizes were normalised by the maximal region size and logarithmized. Second, to further highlight the relationship between the test DSC values and region sizes, we split the area of the instance regions to three equal groups, small, medium and large, based on percentiles. For example, the small group included cases with the region area below the 33th percentile. Boxplots presenting the DSC values for each group for the Yllab method are shown in Fig. 14(b). In this case, there were 99, 53 and 17 cases with DSC values equal to 0 for the small, medium and large group, respectively. Moreover, ANOVA and Tukey's honestly significant difference test indicated that the mean DSC values were significantly different (p-values<0.001) between all three size categories. We repeated the ANOVA analysis and obtained similar findings for the remaining four segmentation networks, indicating that all investigated methods underperformed in the case of smaller regions with respect to the DSC metric.

7. Research impact and future work

The DFU segmentation dataset which has been shared with the research community as part of the DFUC 2022 represents the largest publicly available chronic wound dataset to date. As such, this present challenge report represents the latest insights into the field. Prior works focused on experimenting with smaller datasets at lower image resolutions. Higher resolution DFU wound images present more features and a correspondingly more challenging task. The research works

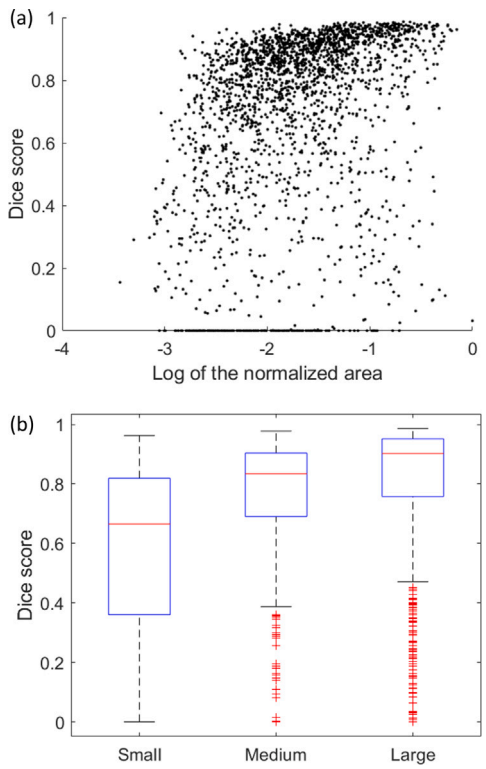

Fig. 14. The relationship (a) between the DSC values and region sizes for the segmentation network from the Yllab team. Additionally, we split (b) the area of the instance regions to three equal groups: small, medium and large. Results indicate that the segmentation network underperformed in the case of the smaller regions in terms of DSC.

detailed in this challenge report provide key insights into the challenges inherent in chronic wound segmentation. The results presented in the present paper highlight the difficulties that underpin the current state of research in the field, particularly in the segmentation of smaller DFU wounds. Chronic wounds can be highly visually complex in nature, especially larger, more developed cases. The range in visual complexity depending on wound development and healing status means that model accuracy is dependent on the model's ability to identify a large range of varied and complex features. In part, such feature variation may be responsible for the challenges in chronic wound segmentation, which may be revealed through the sharing of more diverse datasets and the use of higher resolution images. Future work should focus on a greater understanding of the data used to train segmentation models. A lack of thorough data understanding may be one of the limiting factors in the field. Datasets collected from a single hospital may mean that there are same-patient cases present across training and testing sets, albeit from different visits. Identification of small wounds may prove to be a key facet of fully automated early monitoring systems, which could help patients to seek medical assistance before wounds become more serious. To promote continued research in the field, both the DFUC 2022 dataset and the Grand Challenge platform will continue to be made publicly available after the challenge deadline. It is our intent to conduct further chronic wound related challenges in the near future.

8. Conclusion

DFUC 2022 was conducted to support innovation in computer algorithm development, encourage data sharing, and enable reproducible and multidisciplinary research. Amongst the 26 participating teams, the winning teams set the baselines for this new segmentation dataset, with a DSC of 0.7287. This paper provides an extensive post-challenge analysis. By conducting ensemble methods, we observed marginal performance improvements. The statistical analysis showed that there are no significant differences between the top-2 best performers. We provide further analysis based on region-based segmentation performance, with findings showing a significant positive correlation between the DSC values and DFU region sizes. When we categorised the region sizes to small, medium and large, we found the mean DSC values were significantly different for all categories. Our analysis indicates that the methods proposed by the winning teams underperformed in the segmentation of small DFU regions.

CRediT authorship contribution statement

Moi Hoon Yap: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Writing – original draft, Writing – review & editing. **Bill Cassidy:** Conceptualization, Data curation, Investigation, Writing – original draft, Writing – review & editing. **Michal Byra:** Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Ting-yu Liao:** Methodology, Writing – original draft. **Huahui Yi:** Methodology, Writing – original draft. **Adrian Galdran:** Methodology, Writing – original draft. **Yung-Han Chen:** Methodology, Writing – original draft. **Raphael Brüngel:** Methodology, Writing – original draft. **Sven Koitka:** Methodology, Writing – original draft. **Christoph M. Friedrich:** Methodology, Writing – original draft. **Yu-wen Lo:** Methodology, Writing – original draft. **Ching-hui Yang:** Methodology, Writing – original draft. **Kang Li:** Methodology, Writing – original draft. **Qicheng Lao:** Methodology, Writing – original draft. **Miguel A. González Ballester:** Methodology, Writing – original draft. **Gustavo Carneiro:** Methodology, Writing – original draft. **Yi-Jen Ju:** Methodology, Writing – original draft. **Juinn-Dar Huang:** Methodology, Writing – original draft. **Joseph M. Pappachan:** Data curation, Resources, Supervision, Writing – original draft. **Neil D. Reeves:** Project administration, Supervision, Writing – original draft. **Vishnu Chandrabalan:** Resources, Supervision, Writing – original draft. **Darren Dancey:** Resources, Supervision, Writing – original draft. **Connah Kendrick:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data has been made available for DFUC 2022 participants. It is made available upon request by submitting a licence agreement to the dataset owner.

Acknowledgements

We would like to thank the MICCAI conference for hosting DFUC 2022, and AITIS for sponsoring the wining teams' prizes. We would also like to thank all participants of the challenge for their effort and contributions to DFU research.

References